2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech) | 978-1-7281-6175-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/CloudTech49835.2020.9365916

A Review of Credit Card Fraud Detection Using Machine Learning Techniques

Nadia Boutaher *
Laboratory of Computer Science, Systems
Modeling and Decision Support
Faculty of Sciences AinChock, Hassan II University
Casablanca, Morocco
nadia.boutaher1@gmail.com

Amina Elomri, Noreddine Abghour, Khalid Moussaid and Mohamed Rida
Laboratory of Computer Science, Systems
Modeling and Decision Support
Faculty of Sciences AinChock, Hassan II University
Casablanca, Morocco
{amina.elomri, noreddine.abghour, khalid.moussaid,
mohamed.rida}@univh2c.ma

Abstract—Big Data technologies concern several critical areas such as Healthcare, Finance, Manufacturing, Transport, and E-Commerce. Hence, they play an indispensable role in the financial sector, especially within the banking services which are impacted by the digitalization of services and the evolvement of e-commerce transactions. Therefore, the emergence of the credit card use and the increasing number of fraudsters have generated different issues that concern the banking sector. Unfortunately, these issues obstruct the performance of Fraud Control Systems (Fraud Detection Systems & Fraud Prevention Systems) and abuse the transparency of online payments. Thus, financial institutions aim to secure credit card transactions and allow their customers to use e-banking services safely and efficiently. To reach this goal, they try to develop more relevant fraud detection techniques that can identify more fraudulent transactions and decrease frauds. The purpose of this article is to define the fundamental aspects of fraud detection, the current systems of fraud detection, the issues and challenges of frauds related to the banking sector, and the existing solutions based on machine learning techniques.

Keywords—big data; machine learning techniques; credit card fraud detection; legitimate transaction; financial institutions; digitalization; e-commerce.

I. INTRODUCTION

Nowadays, various economic institutions and financial companies have used big data technologies in their electronic commerce systems to help their customers making online transactions from anywhere and whenever they are. There are different categories of those systems, such as Credit card systems, Telecommunication systems, Insurance systems, and Online auction systems [3] that are used by legitimate users and fraudsters. Unfortunately, more those systems are used more frauds are generated day by day, especially in the banking sector. Then, the fraud Prevention systems (FPSs) are used in order to protect the customers and the companies from these electronic crime issues.

However, with the intelligence of fraudsters and their adaptability to these systems, banks and their customers faced more issues and challenges compared to the past. The fact that leads to use the Fraud detection systems (FDSs) which are more relevant and efficient to detect and recognize fraudulent activities.

Some big data challenges limit the development and effectiveness of Fraud Detection Systems (FDSs) [3]. Then, many researchers have proposed various machine learning techniques that increase low detection accuracy, speed up the time of detection, and reduce false alerts.

This paper tries to focus on the current fraud banking systems, the issues and challenges they faced, and the existing machine learning techniques used to maximize detection rate and reduce the rate of fraud. Also, we have analyzed and compared these techniques to underline their assumptions and objectives.

The remainder of this article provided the following points: First, it defines the theoretical aspects of fraud detection systems. Second, it explains the existing machine learning techniques used to handle banking fraud detection issues and challenges and provides a comparative analysis of these techniques based on different criteria. Third, it defines the perspectives of our paper.

II. THEORETICAL BACKGROUND

A. Anomaly Detection

Anomaly detection is a machine learning method used to identify data points and observations that diverge from the normal behavior of data [8].

It is an issue that has been occurred in various research domains, like intrusion detection, system health monitoring and fraud detection in credit card transactions [9].

In data mining and machine learning fields, we can categorize anomalies as:

- Point anomalies: If a single instance of data deviates from the rest of the data, it is defined as an anomalous point.

- Contextual anomalies: This type of anomaly is identified when a data instance is anomalous in a specific context (but not otherwise), it can be also called a conditional anomaly.

- Collective anomalies: If a group of related data instances is anomalous with respect to each data point of the dataset.

B. *Financial Fraud Detection*

In general sense, financial fraud can be identified when a fraudster takes money or other financial assets from another one through deception or criminal activity [3].

Therefore, cybercriminals and breaches have an effect on the personal information related to customers and the other confidential information stored by financial institutions, which can provide financial losses and a damaged reputation [11].

Financial fraud affects several sectors, especially in the banking sector. Therefore, banking frauds can be classified as below [6][12]:

- Money Laundering: It is an illegal action used to appear the amounts of money generated by criminal activities conform to a legitimate source.

- Mortgage Fraud: It refers to any fraud activity used to obtain a mortgage loan. Especially, it occurs when a homebuyer tries to manipulate significant information in the process of applying for a mortgage loan to purchase a property.

- Credit card fraud: It occurs when fraudsters make illegal operations in credit card transactions. It can be presented in different activities, like losing the credit card or having it stolen, or stealing the confidential information of a credit card (i.e. card-not-present fraud).

C. *Credit Card Fraud Detection*

Credit Card Fraud Detection is one of the important financial fraud detection types because it has generated a huge amount of financial losses more than the past. Several protection mechanisms are used to combat the current credit card frauds, but these methods are not efficient enough to reduce the impact of these frauds such as Fraud Prevention Systems (FPSs).

Therefore, another type of Fraud Detection System (FDS) is identified to be more efficient in detecting fraud in credit card transaction data.

Some many issues and challenges hinder the development and progress of an efficient FDS and lead to many negative effects like high false alerts, slow detection, and low detection accuracy. These issues and challenges are based on Big Data V's characteristics like below:

- Volume challenge: The data in the Credit Card Fraud Detection System (CCFDS) is very huge and the number of attributes that used to define the habit of a cardholder is around 25 attributes on average. So, the need for reduced data is important in terms of reducing the transaction processing time and the complexity in processing a transaction. To fix this issue, various data reduction techniques are used, like dimensionality reduction (PCA) for selecting the most relevant attributes and numerosity reduction approaches for aggregating credit card transactions [3].

- Velocity challenge: This issue occurs when we look to achieve real-time detection by designing an online CCFDS based on big data technologies. For that different algorithms are used to support CCFDS with an accurate online detection technique such as BOAT (Bootstrapped Optimistic Algorithm for Tree Construction) algorithm that reduces training time and Self-Organization Map (SOM) technique that filters the number of transactions [15].

- Veracity challenge: The cardholders change their activities over time and impacted by several reasons. As well as, the fraudster's attacks are evolving, the Credit Card FDS should be more adaptable and efficient to these new types of fraud.

- Value Challenge: Credit card transactions contain much fewer fraudulent instances than legitimate ones, which lead to a skewed class distribution then to imbalanced data. Supervised approaches are the most impacted techniques by this type of issue because the predictions of the training model are too closely fit the majority class (i.e. normal transactions) and ignore the minority class (i.e. fraudulent transactions). Hence, many modeling and training errors are produced and affect Credit Card Fraud Detection System performance.

- Variety Challenge: Data variety refers to various forms of credit card data sources. Therefore, banking institutions might have stored data in several database formats that are characterized by their own unique data model. This variety can be also in the diversity of processes developed by fraudsters to make frauds.

III. Existing techniques

A. *Credit Card Fraud Detection Process*

To build an accurate Credit Card Fraud Detection System, we must follow several stages:

1) Parameterization stage: It presents the initial step that aims to collect the credit card transaction data, loading it into the appropriate database system and preprocessing it into a pre-established format without issues.

2) Training stage: This stage includes the modeling and processing of the fraud detection system. There are different machine learning algorithms and data mining techniques that can be applied, depending on the nature of the data.

3) Detection stage: When the training step is achieved and the model is available, it is evaluated based on different criteria to have the most efficient technique in terms of detection rates.

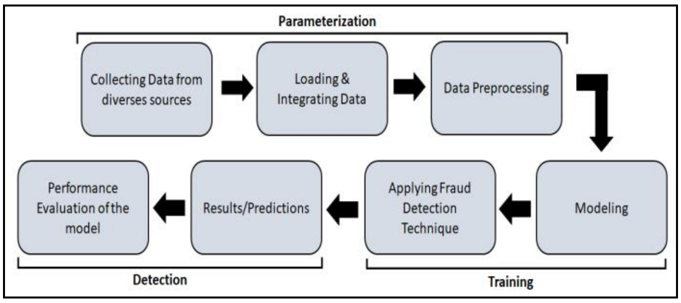
Parameterization
Collecting Data from diverses sources
Loading & Integrating Data
Data Preprocessing
Performance Evaluation of the model
Results/Predictions
Applying Fraud Detection Technique
Modeling
Detection
Training

Figure 1. Process of detecting credit card frauds (Parameterization, Training, and Detection)

To analyze different machine learning approaches that concern the modeling and training stage, we will present various comparative studies between the existing research papers related to the credit card fraud detection area.

B. Theoretical Comparative Study

TABLE I. THEORETICAL COMPARATIVE STUDY

N° of Ref.	Theoretical Criteria		
	Paper Title (Year)	***Purpose / Results***	***Limitations***
[13]	Predictive Modelling For Credit Card Fraud Detection Using Data Analytics (2018)	Random forest provide the best accuracy and precision comparing to the other techniques.	Random Forest lead to the overfitting of trees in memory.
[17]	Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection (2019)	Unsupervised techniques solve the dataset skewness and provide best results than other techniques like Isolation Forest and Local Outlier Factor.	The work should be focusing on resampling techniques that reduce high imbalance rate.
[14]	Credit Card Fraud Detection - Machine Learning methods (2019)	Random Forest algorithm perform efficiently by applying feature selection PCA (Principal Component Analysis) and oversampling SMOTE (Synthetic Minority Over-sampling) techniques.	If there is not much data, it is more relevant to work with classical algorithms but if we have more data that can make an issue, other advanced techniques will be needed.
[16]	Credit Card Fraud Detection Using Convolutional Neural Networks (2016)	The Convolutional Neural Network (CNN) model achieves the best performance.	the proposed method has best results but the future methods should focus on handling the issue of highly imbalanced data.
[18]	Detecting Credit Card Fraud Using Selected Machine Learning Algorithms (2019)	In this work, they used 2 approaches to detect credit card transaction frauds: statistic and incremental learning that shows the best results.	The statistic learning approach cannot be considered as a long-term solution especially for the dataset used.
[19]	Horse Race Analysis in Credit Card Fraud—Deep Learning, Logistic Regression, and Gradient Boosted Tree (2017)	After comparing the predictive performance of 3 machine learning algorithms, we observe that Neural Network (NN) is the best one.	There is some limitations such as the less predictive power of Logistic regression, the large volume of data for GBT and the feature selection issue for NN.
[21]	An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies (2018)	Logistic Regression and Decision Tree have the most accurate results.	Hence, the machine learning models used in this work ignore the other performance metrics.
[22]	A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance (2018)	This paper compare between 10 machine learning algorithms without and with the "Time" feature to capture the performance differences.	Hence, the work should focus on a more efficient way to handle the high level of the data imbalance problem.

C. Comparative Study based on Big Data Challenges

TABLE II. BIG DATA COMPARATIVE STUDY

N° of Ref.	Big Data Challenges			
	Volume of Data		***Speed of Data***	***Imbalance of Data***
	Features	***Transactions***		
[13]	20 attributes (7 are numerical / 13 categorical)	1000 transactions (30% frauds & 70% normals)	Real Time	None
[14]	31 numerical features	284807 transactions (0.173% frauds & 99.827% normals)	Batch	Highly Imbalanced
[17]	28 features	284807 transactions (0.173% frauds & 99.827% normals)	Batch	Highly Imbalanced
[16]	NA	260 million transactions (4000 fraudulent transactions)	Real Time	Highly Imbalanced
[18]	28 numerical PCA variable, time, amount and class	284807 transactions (0.173% frauds & 99.827% normals)	Real Time	Highly Imbalanced
[19]	69 attributes	80 million account level transactions (99.864% legitimate & 0.136% frauds)	Batch	Highly Imbalanced
[21]	28 features	284315 legitimate records in data whereas 492 fraudulent records	Batch	Highly Imbalanced
[22]	28 features	284807 transactions (0.173% frauds & 99.827% normals)	Batch	Highly Imbalanced

D. *Comparative Study of Existing Machine Learning Techniques based on Metrics of Performance*

TABLE III. COMPARISON OF EXISTING MACHINE LEARNING TECHNIQUES

N° of Ref.	Machine Learning Techniques	Metrics of Performance			
		Accuracy (%)	***Precision (%)***	***Fallout (FPR %)***	***Miss Rate (FNR %)***
[13]	Logistic Regression	72	76	64	12
	Decision Tree	72	89	44	25
	Random Forest	76	93	31	23
[14]	Logistic Regression	97.46	58.82	3	8
	Random Forest	99.96	96.38	0	18
	Naive Bayes	99.23	16.17	0.7	17
	Multilayer Perceptron	99.93	79.21	0	18
[17]	Naive Bayes	89	6	NA	NA
	Random Forest	95	99	NA	NA
	SVM	93	NA	NA	NA
	Artificial Neural Network	88	99	NA	NA
	Logistic Regression	94	99	NA	NA
	Isolation Forest	99	99	NA	NA
	Local Outlier Factor	99	99	NA	NA
	K-Means	50	99	NA	NA
[16]	Neural Network	***F1 Score***		0.29	
	SVM			0.27	
	Random Forest			0.30	
	CNN			0.32	
[18]		***Static Learning***		***Incremental Learning***	
		AUC	***AP***	***AUC***	***AP***
	Random Forest	0.91	0.85	0.90	0.83
	SVM	0.89	0.80	0.87	0.80
	Logistic Regression	0.91	0.73	0.91	0.84
[19]	Logistic Regression	***AUC Values***		0.82	
	Gradient Boosted Trees			0.86	
	Deep Learning (NN)			0.88	
[21]	Logistic Regression	99.75	NA	NA	NA
	Decision Tree	99.21	NA	NA	NA
	Xtream Gradient Boosting	98.8	NA	NA	NA
[22]	Logistic Regression	94[a] / 94[b]	95[a] / 94[b]	NA	NA
	Decision Tree	90[a] / 90[b]	91[a] / 89[b]	NA	NA
	Naïve Bayes	91[a] / 91[b]	93[a] / 91[b]	NA	NA
	SVM	94[a] / 94[b]	94[a] / 92[b]	NA	NA
	KNN	93[a] / 93[b]	94[a] / 93[b]	NA	NA
	Random Forest	93[a] / 95[b]	94[a] / 94[b]	NA	NA
	ADB (Adaptive Boost)	94[a] / 95[b]	94[a] / 95[b]	NA	NA
	BAG (Bagging)	93[a] / 94[b]	93[a] / 95[b]	NA	NA

a. Accuracy/Precision (%) With "Time" Feature.

b. Accuracy/Precision (%) Without "Time" Feature.

E. *Synthesis and discussion*

To identify the most relevant machine learning techniques that handle and solve the issues of credit card fraud detection, we analyzed and compared current research contributions using various criteria.

As a purpose, in the current literature, the most used machine learning techniques are supervised techniques like Logistic Regression that always has high accuracy and works well with linear data, SVM method that reduce the time of detection, Random Forest and Neural Network whiches improve the classification rates (FNR and FPR).

These comparative studies help to identify several issues related to credit card fraud detection area such as:

- A lack of publicly available datasets: The majority of uses cases in this comparative study are based on one specific data set from the Kaggle platform.
- Large Datasets: Credit Card Fraud Detection System works fast and extremely efficient with an online big data.
- The real-time processing is not implemented in the most of machine learning techniques: When the model is applied in a real-time context, the detection becomes easier, but the credit card fraud detection model cannot be adaptable to any new attack of fraudsters.
- The highly imbalanced data: Researchers have used several techniques to handle the skewed class like oversampling and undersampling that make the detection system more accurate.

IV. CONCLUSION & FUTURE WORK

Credit Card Fraud Detection highly impacts the financial area. To avoid fraud losses, bank institutions tried to develop new and advanced fraud detection techniques. Hence, this paper aimed to present the theoretical aspects of the credit card detection issue and provided different comparative studies of the existing machine learning techniques used to deal with credit card frauds.

It is concluded from this analysis that supervised algorithms are much used than other types of techniques like Logistic Regression, Random Forest, and SVM.

As perspectives, our work will focus on the relevant machine learning algorithms identified in this paper and analyze data engineering techniques, in order to propose our contribution that can handle the imbalanced data, make the fraud detection system more relevant to the real-time problem, and provide more efficient classification metrics.

ACKNOWLEDGMENT

We thank all members of Laboratory of Computer Science, Systems Modeling and Decision Support for their suggestions and help.

REFERENCES

[1] I. Benchaji and S. Douzi, "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection," p. 5.

[2] L. Nahar, I. Amir, and S. Shabnam, "A Comprehensive Survey of Fraud Detection Techniques," Int. J. Appl. Inf. Syst., vol. 10, no. 2, pp. 26–32, Dec. 2015.

[3] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," J. Netw. Comput. Appl., vol. 68, pp. 90–113, Jun. 2016.

[4] A. Kundu, S. Sural, and A. K. Majumdar, "Two-Stage Credit Card Fraud Detection Using Sequence Alignment," in Information Systems Security, vol. 4332, A. Bagchi and V. Atluri, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 260–275.

[5] K. T. Hafiz, S. Aghili, and P. Zavarsky, "The use of predictive analytics technology to detect credit card fraud in Canada," in 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), Gran Canaria, Spain, 2016, pp. 1–6.

[6] R. R. Popat and J. Chaudhary, "A Survey on Credit Card Fraud Detection Using Machine Learning," in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, 2018, pp. 1120–1125.

[7] J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," Comput. Secur., vol. 57, pp. 47–66, Mar. 2016.

[8] S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," Procedia Comput. Sci., vol. 60, pp. 708–713, 2015.

[9] M. Ahmed, A. N. Mahmood, and Md. R. Islam, "A survey of anomaly detection techniques in financial domain," Future Gener. Comput. Syst., vol. 55, pp. 278–288, Feb. 2016.

[10] A. Alnafessah and G. Casale, "Artificial neural networks based techniques for anomaly detection in Apache Spark," Clust. Comput., Oct. 2019.

[11] "Preventing and Detecting Financial Institution Fraud — ACFE Insights.html." .

[12] I. Sadgali, N. Sael, and F. Benabbou, "Performance of machine learning techniques in the detection of financial frauds," Procedia Comput. Sci., vol. 148, pp. 45–54, 2019.

[13] S. Patil, V. Nemade, and P. K. Soni, "Predictive Modelling For Credit Card Fraud Detection Using Data Analytics," Procedia Comput. Sci., vol. 132, pp. 385–395, 2018.

[14] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," in 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2019, pp. 1–5.

[15] Y. Dai, J. Yan, X. Tang, H. Zhao, and M. Guo, "Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies," in 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, 2016, pp. 1644–1651.

[16] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks," in Neural Information Processing, vol. 9949, A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, and D. Liu, Eds. Cham: Springer International Publishing, 2016, pp. 483–490.

[17] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection," in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 320–324.

[18] M. Puh and L. Brkic, "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms," in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1250–1255.

[19] G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, "Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree," in 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2017, pp. 117–121.

[20] L. Zheng et al., "A new credit card fraud detecting method based on behavior certificate," in 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, 2018, pp. 1–6.

[21] T. Choudhury, G. Dangi, T. P. Singh, A. Chauhan, and A. Aggarwal, "An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies," in 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 591–597.

[22] S. Rajora et al., "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 2018, pp. 1958–1963.